STA108 Final Project

Manish Rathor

6/12/2023

```
# I will use these packages to calculate/visualize relevant values
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
getwd()
```

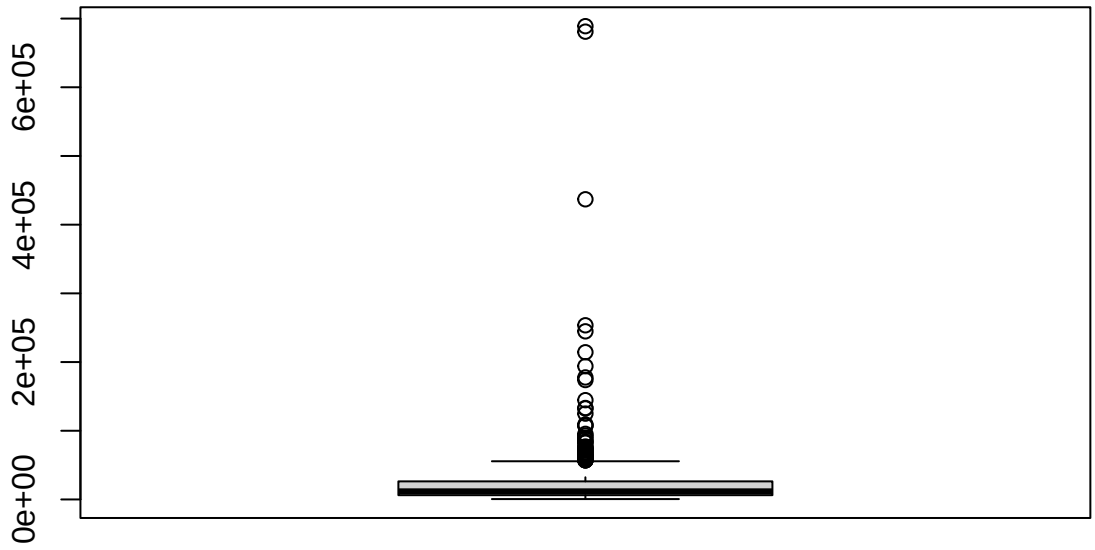
```
## [1] "/Users/manishrathor/Documents/School /Davis /2022-23/SQ2023/STA108 "
```

```
#reading the data into RStudio and naming the columns
demographic <- read.table("/Users/manishrathor/Documents/School /Davis /2022-23/SQ2023/STA108 /Data/dem
names(demographic) <- c("ID", "County", "State", "LandArea", "TotalPopulation", "PercPop18-34", "PercPo
head(demographic)
```

```
##   ID      County State LandArea TotalPopulation PercPop18-34 PercPop65+
## 1  1 Los_Angeles    CA     4060         8863164         32.1        9.7
## 2  2        Cook    IL      946         5105067         29.2       12.4
## 3  3      Harris    TX     1729         2818199         31.3        7.1
## 4  4   San_Diego    CA     4205         2498016         33.5       10.9
## 5  5      Orange    CA      790         2410556         32.6        9.2
## 6  6       Kings    NY       71         2300664         28.3       12.4
##   Physicians HospitalBeds SeriousCrimes PercentHSGrads PercentBachDegree
```

```
## 1      23677        27700        688936           70.0             22.3
## 2      15153        21550        436936           73.4             22.8
## 3       7553        12449        253526           74.9             25.4
## 4       5905         6179        173821           81.9             25.3
## 5       6062         6369        144524           81.2             27.8
## 6       4861         8942        680966           63.7             16.6
##   PercentBelowPoverty PercentUnemployment PerCapitaIncome TotalPersonalIncome
## 1                11.6                 8.0           20786              184230
## 2                11.1                 7.2           21729              110928
## 3                12.5                 5.7           19517               55003
## 4                 8.1                 6.1           19588               48931
## 5                 5.2                 4.8           24400               58818
## 6                19.5                 9.5           16803               38658
##   GeoRegion
## 1         4
## 2         2
## 3         3
## 4         4
## 5         4
## 6         1
```

```
crimedata <- demographic[,10]
boxplot(crimedata)
```



Part 1: Creating the Regression Model

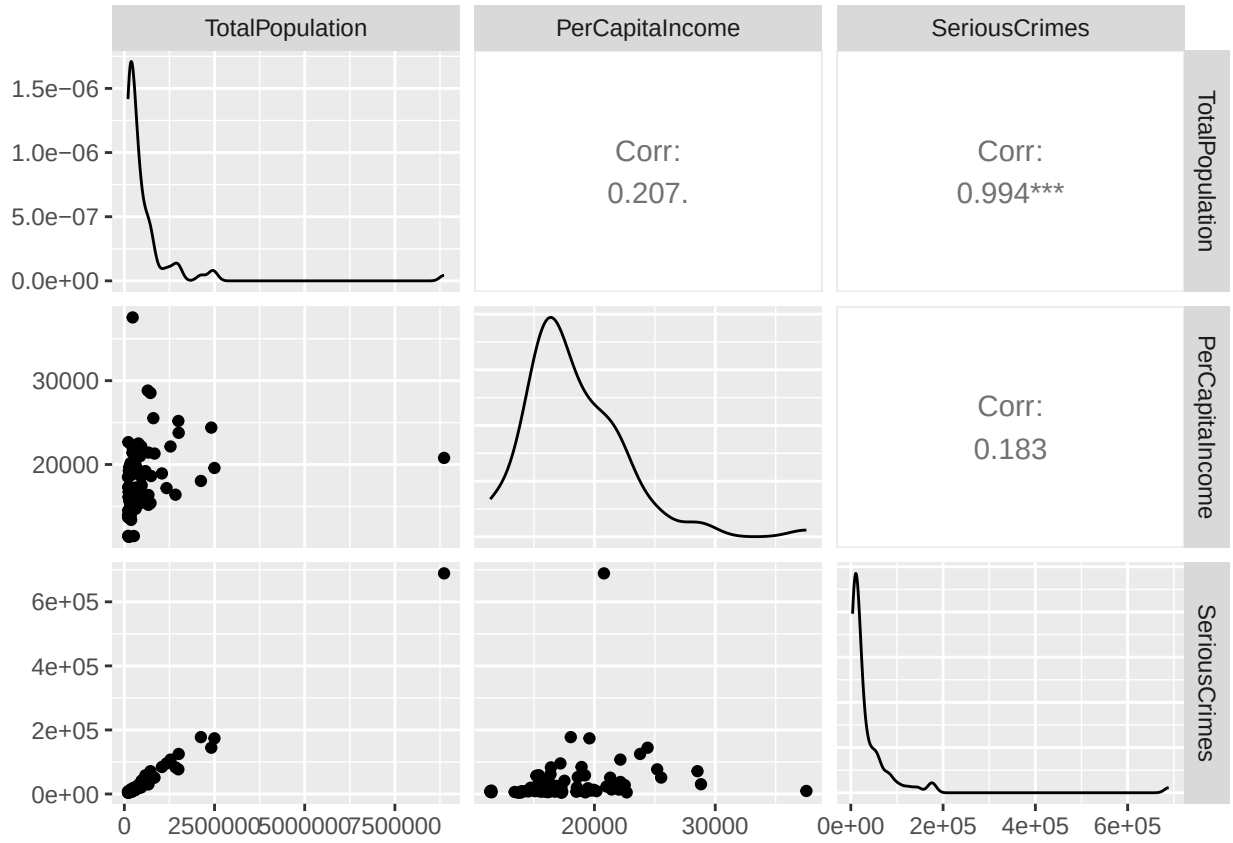
```
#creating a new data frame that only contains counties in geographic region 4
demographic4 <- filter(demographic, GeoRegion == "4")
head(demographic4)
```

```
##   ID      County State LandArea TotalPopulation PercPop18-34 PercPop65+
## 1  1 Los_Angeles    CA     4060         8863164         32.1        9.7
## 2  4   San_Diego    CA     4205         2498016         33.5       10.9
## 3  5      Orange    CA      790         2410556         32.6        9.2
## 4  7    Maricopa    AZ     9204         2122101         29.2       12.5
## 5 12        King    WA     2126         1507319         30.1       11.1
## 6 13 Santa_Clara    CA     1291         1497577         32.6        8.7
##   Physicians HospitalBeds SeriousCrimes PercentHSGrads PercentBachDegree
## 1      23677        27700        688936           70.0              22.3
## 2       5905         6179        173821           81.9              25.3
## 3       6062         6369        144524           81.2              27.8
## 4       4320         6104        177593           81.5              22.1
## 5       5280         4009        124959           88.2              32.8
## 6       4101         3342         77009           82.0              32.6
##   PercentBelowPoverty PercentUnemployment PerCapitaIncome TotalPersonalIncome
## 1                11.6                 8.0           20786              184230
## 2                 8.1                 6.1           19588               48931
## 3                 5.2                 4.8           24400               58818
## 4                 8.8                 4.9           18042               38287
## 5                 5.0                 4.6           23779               35843
## 6                 5.0                 5.5           25193               37728
##   GeoRegion
## 1         4
## 2         4
## 3         4
## 4         4
## 5         4
## 6         4
```

```
#creating a data frame that contains only the variables used in the regression model
crimemodeldata <- demographic4[,c(5,15,10)]
head(crimemodeldata)
```

```
##   TotalPopulation PerCapitaIncome SeriousCrimes
## 1         8863164           20786        688936
## 2         2498016           19588        173821
## 3         2410556           24400        144524
## 4         2122101           18042        177593
## 5         1507319           23779        124959
## 6         1497577           25193         77009
```

```
#visualizing/calculating the correlation between the variables
ggpairs(crimemodeldata)
```



I chose the Total Population and Per Capita Income as the predictor variables because they are relatively highly correlated with the Total Number of Serious Crimes (Y).

Neither of the predictor variables are highly correlated with each other, so we do not need to be concerned about multicollinearity.

```
crimemodel <- lm(SeriousCrimes ~ TotalPopulation + PerCapitaIncome, crimemodeldata)
crimemodel
```

```
##
## Call:
## lm(formula = SeriousCrimes ~ TotalPopulation + PerCapitaIncome,
##     data = crimemodeldata)
##
## Coefficients:
##     (Intercept)  TotalPopulation  PerCapitaIncome
##      3804.83220          0.07741         -0.46867
```

The regression model is:

$$\hat{Y} = 3804.83220 + 0.07741X_1 - 0.46867X_2$$

β_0 is the y-intercept.

β_1 represents the rate of change of the predicted number of total serious crimes based on the change in the total population (when the per capita income is held constant).

β_2 represents the rate of change of the predicted number of total serious crimes based on the change in per capita income (when the total population is held constant).

Part 2: 90% Confidence Interval for B_j (j = 1,2)

```
#This function allows us to create confidence intervals for each parameter in the model
confint(crimemodel, level = 0.9)
```

```
##                             5 %          95 %
## (Intercept)     -4.008474e+03  1.161814e+04
## TotalPopulation  7.573344e-02  7.907676e-02
## PerCapitaIncome -8.918714e-01 -4.547556e-02
```

We are 90% confident that the true value of β_1 lies within the interval $(0.07573344, 0.07907676)$.

We are 90% confident that the true value of β_2 lies within the interval $(-0.8918714, -0.04547556)$.

Part 3: P-Value of Hypothesis Test for Regression Relation at alpha = 0.01

Alternatives

H_0: There is no regression relation. $\beta_1 = \beta_2 = 0$

H_a: There is a regression relation. Not all $\beta_j = 0$, $(j = 1, 2)$

Decision Rule

If the p-value $> \alpha$, conclude H_0

If the p-value $< \alpha$, conclude H_a

Conclusion

```
#We can use the ANOVA table to find the p-value
anova(crimemodel)
```

```
## Analysis of Variance Table
##
## Response: SeriousCrimes
##                  Df     Sum Sq    Mean Sq   F value  Pr(>F)
## TotalPopulation   1 5.2842e+11 5.2842e+11 6153.4531 < 2e-16 ***
## PerCapitaIncome   1 2.9222e+08 2.9222e+08    3.4029 0.06908 .
## Residuals        74 6.3547e+09 8.5874e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value $= (2 \times 10^{-16}) + (0.06908) = 0.06908$

$0.06908 > 0.01$. Because the p-value $> \alpha$, we cannot reject the null hypothesis and must conclude H_0. At the $\alpha = 0.01$ level, there is no regression relation, therefore all $\beta_j = 0$. However, it is important to note that for F-tests, multiple predictor variables are included in the calculations. Because the p-value does not account for interaction effects between predictor variables, it should not be used as the sole value to make conclusions.

Part 4: Hypothesis Test for Regression Relation at alpha = 0.05 + Adjusted R-Squared

To test for regression relation, we will use an F-Test because it allows us to account for the interaction effects among the predictor variables.

Alternatives

H_0: There is no regression relation. $\beta_1 = \beta_2 = 0$

H_a: There is a regression relation. Not all $\beta_j = 0$ $(j = 1, 2)$

Decision Rule

If $F* \leq F(1-\alpha, p-1, n-p)$, conclude H_0

If $F* > F(1-\alpha, p-1, n-p)$, conclude H_a

In the F critical value, α represents the significance level, p represents the number of variables in the model, and n represents the number of observations used to create the model.

In this case, $\alpha = 0.05$, $p = 3$, and $n = 77$.

Conclusion

```
#We can use the summary function to find the F statistic (F*) and the adjusted R^2
summary(crimemodel)
```

```
##
## Call:
## lm(formula = SeriousCrimes ~ TotalPopulation + PerCapitaIncome,
##     data = crimemodeldata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34435  -2813   -111   3924  24763
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.805e+03  4.691e+03   0.811   0.4199
## TotalPopulation  7.741e-02  1.004e-03  77.130   <2e-16 ***
## PerCapitaIncome -4.687e-01  2.541e-01  -1.845   0.0691 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9267 on 74 degrees of freedom
## Multiple R-squared:  0.9881, Adjusted R-squared:  0.9878
## F-statistic:  3078 on 2 and 74 DF,  p-value: < 2.2e-16
```

$F* = 3078$

```
#We can use this function to find the F critical value
qf(0.95,2,74)
```

```
## [1] 3.120349
```

$F(1-0.05, 3-1, 77-3) = F(0.95, 2, 74) = 3.120349$

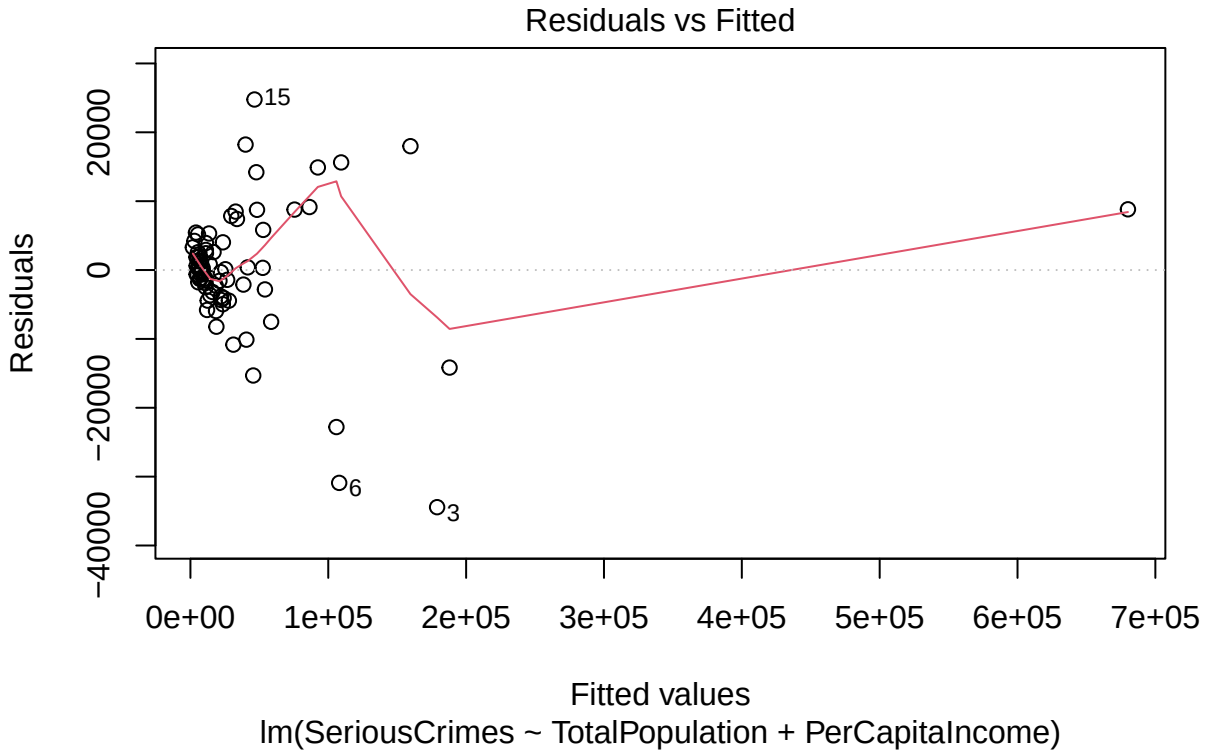
$3078 > 3.120349$. Because $F* > F(1-\alpha, p-1, n-p)$, we can reject the null hypothesis and conclude H_a. This means there is a regression relation and not all β_j is equal to 0.

The R^2 value is a value that measures the goodness of fit of a regression model.It indicates the proportion of the variation in the response variable that is explained by the predictor variables. The adjusted R^2 value does the same thing, but it accounts for the number of predictor variables as well as the sample size of the data. It does this by avoiding the inclusion of statistically insignificant predictors in its calculations. For this model, the adjusted $R^2 = 0.9878$. This means that 98.78% of the variation in the response variable can be explained by the predictor variables. This is a high value, meaning that there is very little unexplained variation in the response data.

Part 5: Diagnostic Plots

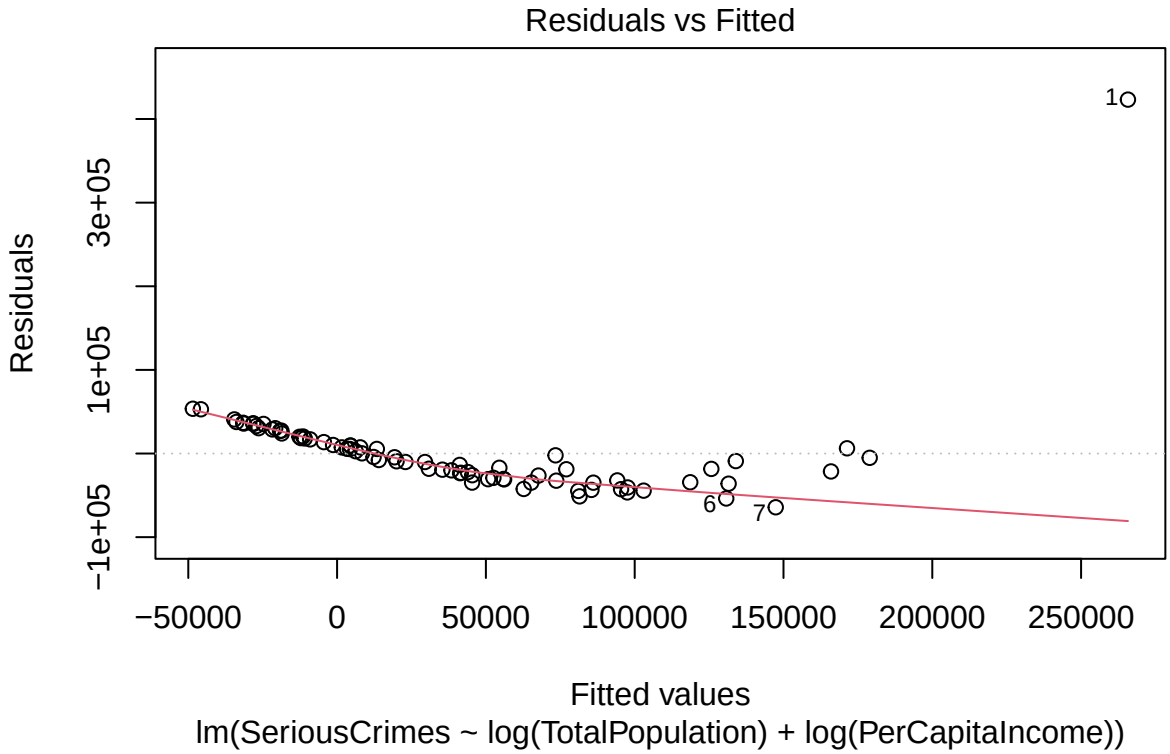
```
residuals <- crimemodel$residuals
```

```
plot(crimemodel, 1)
```



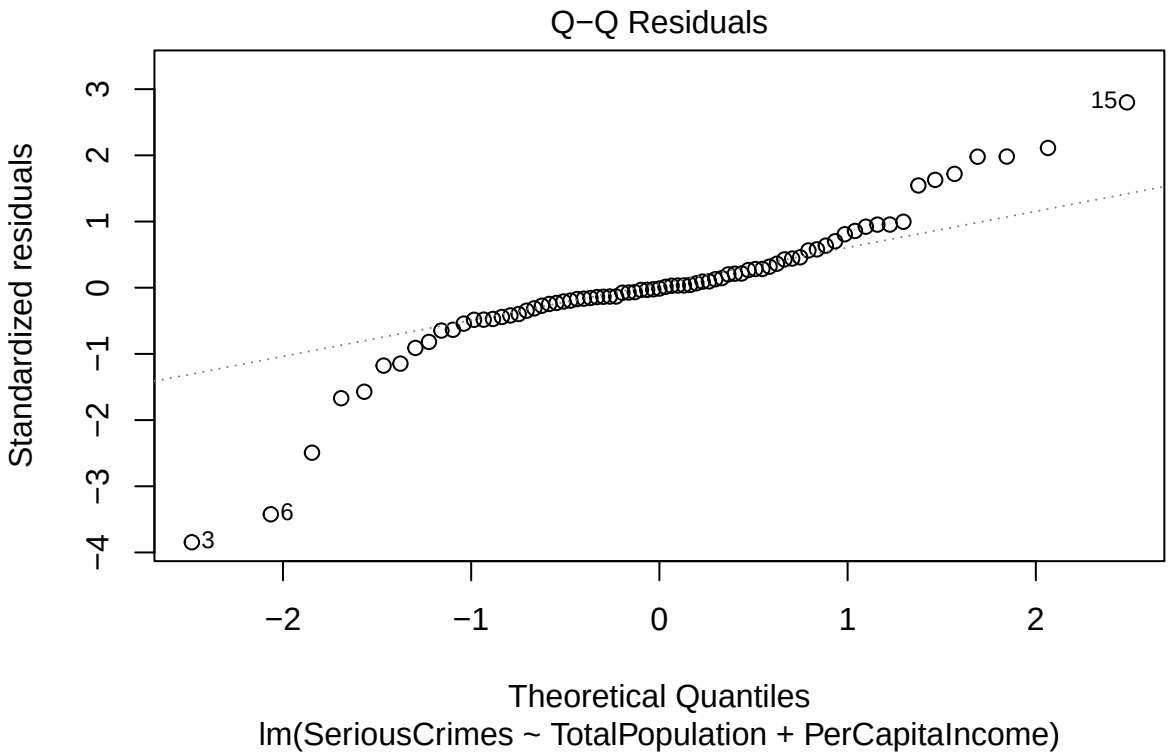
The Residuals vs. Fitted plot is used to check the linear relationship assumption. A horizontal red line is an indicator that the data is linearly related. However, we can see that the red line is not horizontal, and the residuals are clumped together. This means that the relationship may not be linear. Applying a non-linear transformation to the predictors may fix this issue.

```
crimemodel2 <- lm(SeriousCrimes ~ log(TotalPopulation) + log(PerCapitaIncome), crimemodeldata)
plot(crimemodel2, 1)
```



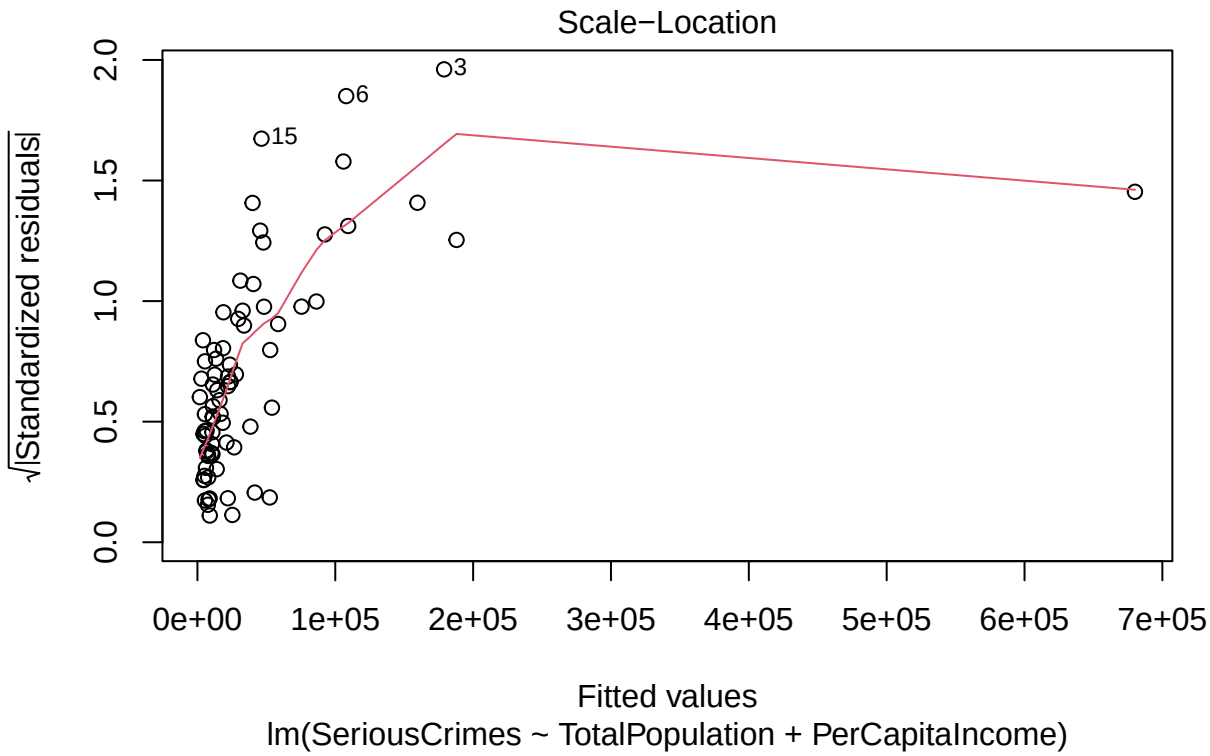
As we can see, by applying the transformation "log(x)", the residuals show a much more linear relationship.

```
plot(crimemodel, 2)
```



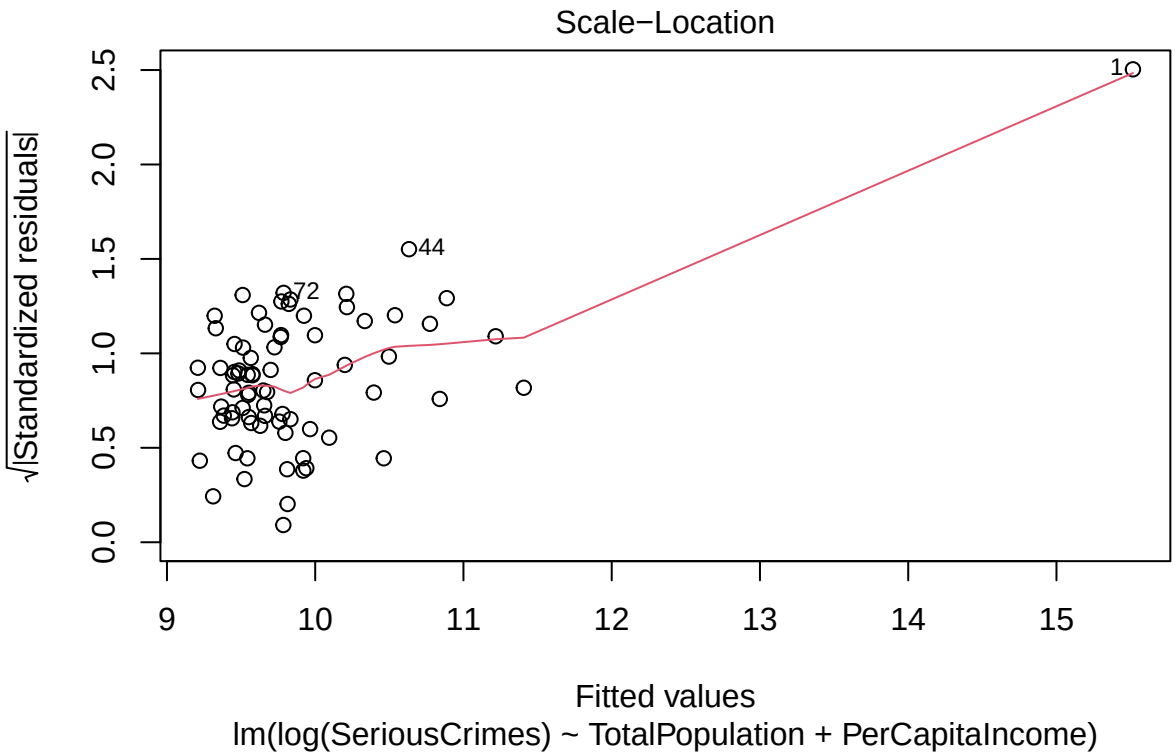
The normal Q-Q plot is used to examine whether the residuals are normally distributed. If the residuals remain on the dashed line ($y = x$), we can conclude that they are normally distributed. We can see that the majority of the residuals lie on the line. However, there are some outliers that do not lie on the plane.

```
plot(crimemodel, 3)
```



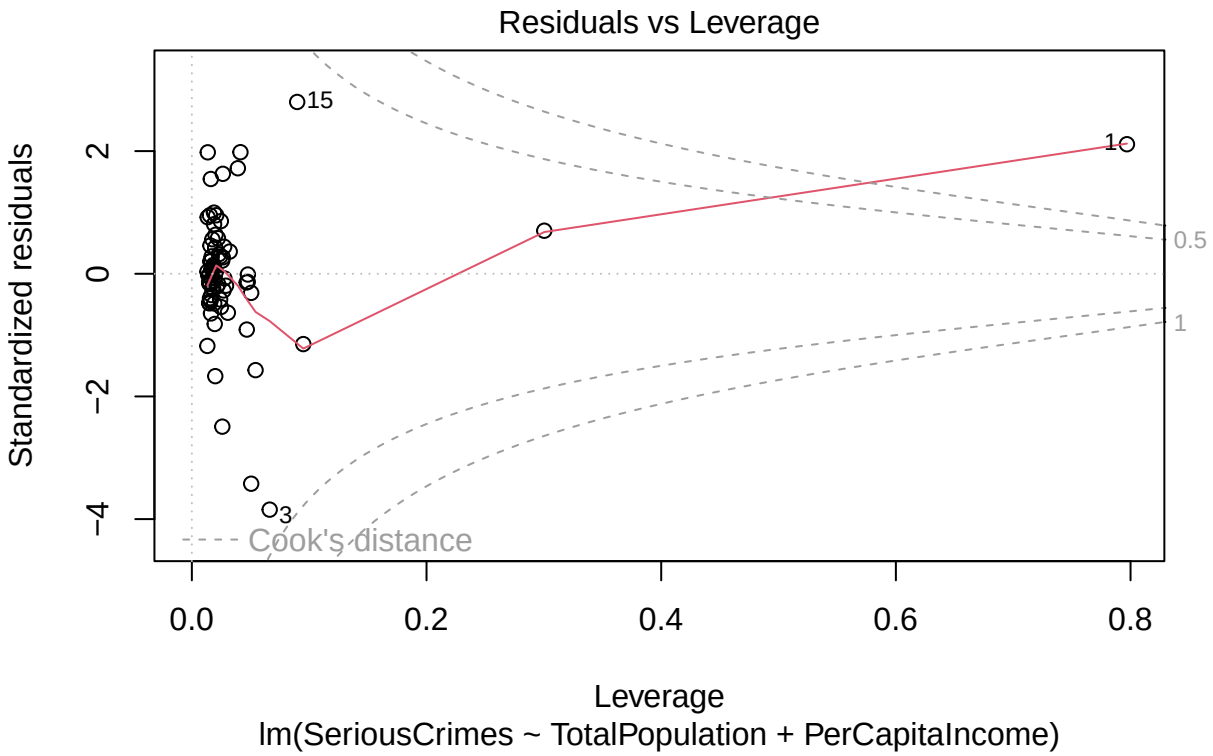
The Scale-Location plot is used to check the homogeneity of the variance. The variance can be considered homogeneous if the points are equally spread out and the red line is approximately horizontal. However, this is not the case in our plot. To try and fix this, we can apply a non-linear transformation to the response variable.

```
crimemodel3 <- lm(log(SeriousCrimes) ~ TotalPopulation + PerCapitaIncome, crimemodeldata)
plot(crimemodel3, 3)
```



The application of a log(y) transformation creates more spread among the points and causes the red line to become more linear. However, the data points are still clumped together, and the red line is not horizontal. We will have to conclude that the variance is not homogeneous.

```
plot(crimemodel, 5)
```



The Residuals vs. Leverage plot checks for outliers. We can see that there are several points that can be considered outliers. Point 1 is an extreme outlier, as it lies outside Cook's Distance. This means that it can be considered a high leverage point, and will skew the results of the regression analysis with its inclusion or exclusion. Therefore, we do not want to simply remove it.